相澤裕介『統計処理に使うExcel2010活用法』

カットシステム

# 理論的な要点（第1章）

文責：柳瀬陽介（2010/11/30）

以下、教科書とウェブ配布する「ファイル」のデータを使いながら、統計処理の基本を学んでゆきましょう。⇒は課題を示すサインですので、必ずエクセルを動かしながら課題を実行してください。

(a) 平均値だけではデータのばらつき具合はわからない。

だから、メディアン（中央値 median）とモード（最頻値 mode）についても学ぼう。

例えば10組のカップルが結婚式に使った費用に関するデータ{200,250,300,320,320,340,400,430,550,10000; 単位は万円}に関してはどの代表値を使えばいいのか具体的に説明しなさい

また、データの最大値(MAX)、最小値(MIN)を示すことも通常求められている。

⇒エクセルの関数機能を使って、ファイルのそれぞれのテスト結果の、AVERAGE, MEDIAN, MODE, MAX, MINを求めなさい。（pp. 8-16）

注：離れたセル範囲の値を関数で求めるには、セル範囲をカンマ(,)で区切る。例えば「B2:B31」「D62:D31」という別々の行に別れたセル範囲のデータ全体の平均値を一括に出すには「=AVERAGE(B2:B31,D2:D31)」のように、二つのセル範囲をカンマ(,)で区切って入力する。

あるいは関数を「数式」→「関数の挿入」で選択した場合には、ダイアログボックスに「数値1」「数値2」というデータ入力欄が出てくるので、そこにそれぞれの範囲を入力する。

なおたいていの場合、左クリックをしたままマウスを動かしてセルの範囲指定をすると、その指定された範囲がそのまま入力データ範囲となるので、適当にマウスを動かして、各自にとって便利な方法を見つけること。

(b) データのばらつき具合を直感的に表すには？　⇒　「ヒストグラム」

⇒ファイルのデータでヒストグラムを作りなさい。(pp.16-24を参照すること。ただし適宜値などは変えて自分が目的とするヒストグラムを作ること。ただしヒストグラムの場合は、グラフの棒と棒の間を詰めることが作法とされていますので、そのようにヒストグラムを作りかえて下さい)

(c) データのばらつき具合を数学的に表すには？　⇒　「分散」(variation)

(d) 分散の表現としてはどれが数学的に合理的？またそれは何故？

(1)　Σ（値　-　平均値）　　　←これではなぜ駄目？

(2)　Σ （値　–　平均値）2 ←これではなぜ駄目？

(3)　$\frac{Σ （値　–　平均値）^{2}　}{データの個数　}$　　　←【分散の定義】

　分散のエクセル関数は？ (VARP)

PはPopulation (母集団) あるいはParameter（母数：母集団に関する代表値）を示すPと覚えておこう！

注：Excel 2010から関数はVAR.Pとなりましたが、従来のVARPも用意されているはずです。

⇒ファイルのデータで、それぞれの分散を求めなさい（pp. 29-34）

(e) 分散よりも実は標準偏差 (Standard Deviation) の方がよく計算では使われる。それはなぜ？

(4)　標準偏差　=　 $\sqrt{分散}$

=　$\frac{\sqrt{Σ（値-平均値）^{2}}}{データの個数}$ ←【標準偏差の定義】

※上ではうまく表記できていませんが、ルートは「データの個数」にももちろんかかっています！

他のサイトから表記をコピーしますと、以下のようになります。



<http://www.okasan-online.co.jp/ont/use/stepup/tec_basic/tec_basic_03.html>

標準偏差のエクセル関数は？　（STDEVP）

注：Excel 2010から関数はSTDEV.Pとなりましたが、従来のSTDEVPも用意されているはずです。

⇒ファイルのデータで、それぞれの標準偏差を求めなさい（pp. 35-37）

(f) 分散・標準偏差を使えば、異なる種類のデータ（例、身長と体重）が、それぞれの母集団の中でどのような位置づけにあるかを数学的に表現できる（z得点）。

(5)　$\frac{値-平均値}{標準偏差}$　　　　　　←　【z得点の定義】

例えばA君は、身長は175センチで、体重は70キロである。身長と体重の全国平均が仮に165センチ、60キロとすると、A君はそれぞれ「10」大きいから、身長と体重において同じ程度に平均から離れていると言えるだろうか。この計算のいびつさを補正するにはどうしたらよいか考えてみよう。

⇒ファイルのデータで、一つ一つのデータのz得点を求めなさい。（数式でz得点を定義して、あとはその数式をコピーすること。この際、「絶対参照」が必要になります）。

「絶対参照」を理解するために、その対概念である「相対参照」を理解しましょう。

エクセルでは、あるセルに数式を入れて、そのセルをコピーして貼りつけると、エクセルがうまく数式を自動的に変えて同じ計算結果が出てくる用にしてくれます。例えば、あるC1のセルに「＝A1+B1」という数式を入れ、これをコピーしてC2からC10に貼り付けます。そうしますとC2には「＝A2+B2」、C3には「＝A3+B3」、C4には「＝A4+B4」というように自動的に計算式が調整されて入力されその結果が表示されています（初めてこれを知った時はちょっと感動ものです）。このようにコピーする時に、関連セルを「相対的に」[＝周りの関係に応じて]参照してくれるのを「相対参照」といいます。

しかし「相対参照」ではかえって困る場合もあります。例えば、あるセル（A12）にA列（A1からA10までのセル）の平均値が入れられているとしましょう。そして計算は、そのA列の平均値（A12の値）と、B列のそれぞれのセル（B1からB10まで）を足すものだとしましょう。つまりD1の計算式は「＝A12+B1」となります。ですが、このD1をコピーして貼り付けますと、相対参照によりD2の計算式は「＝A13+B2」となってしまいます。エクセルが気を効かせてB列と同じようにA列も一つずつセルを下げたわけです。しかしA13は空欄ですから困ります。A12は固定して参照してもらわなければならないのです。

ここで絶対参照の登場です。絶対参照はあるセルを固定します。やり方は簡単で、固定したいセルの列記号と行番号の前にそれぞれ「$」マークをつけます。「A12」を固定するには「$A$12」と入力します。これが絶対参照です。

D1の式を「＝$A$12+B1」としておいて、これをD2以下に貼りつけると、D2は「＝$A$12+B2」、D3は「＝$A$12+B3」と、$マークで固定した絶対参照は変わらず、$マークをつけない相対参照は自動的に位置が調整されて計算式が生成されます。この絶対参照と相対参照の使い分けは必須知識として覚えておいてください。

ちなみに行と列の一方を変化、他方を固定にする参照法は「複合型」と呼ばれます。固定したいどちらか（つまり行か列のどちらか）に$マークをつけます。つまりA列ということは固定して、行番号だけを自動調整したいなら、たとえば「$A1」と入力します。ついでに覚えておいてください。

(g) しかしz得点は、-1.2, 0.2, 1.6などといった数字で表現されるため、直感的にわかりやすい。これを日常的にわかりやすい数字に加工したのが偏差値である。

(6)　$ （\frac{値-平均値}{標準偏差}　x　10）+50$

$=（z得点　x　10）+50$　　←　【偏差値の定義】

z得点を偏差値に補正することにより、どのような直観的な理解が得られるだろうか。

覚えておこう！

偏差値40-60の間に全データの約68%が含まれる。

偏差値30-70の間に全データの約95%が含まれる。

偏差値20-80の間に全データの約99%が含まれる。

⇒ファイルのデータで、一つ一つのデータの偏差値を求めなさい。（数式で偏差値を定義して、あとはその数式をコピーすること。この際、「絶対参照」が必要になります。pp. 34-44）

エクセルではz 得点をSTANDARDIZE関数により求めることもできます。関数の書式は次の通りです。

=STANDARDIZE(素得点,平均,標準偏差)

平均を求めるにはAVERAGE関数、標準偏差を求めるには（STDEVか）STDEVP関数を使いますから、Excelで偏差値を計算したいなら次のように入力してください。

=STANDARDIZE(素得点,AVERAGE(素得点の範囲),STDEVP(素得点の範囲))\*10+50

もちろん「素得点の範囲」で指定せずに、平均と標準偏差をいったんあるセルに算出しておき、そのセルを絶対参照するやり方もできます。というより、その方が関数式がシンプルになるので、わかりやすいでしょう（関数式をあまり複雑にすると間違いやすくなります）。

※考えてみよう

・ある地方ののんびりした高校で勉強していたA君は、校内実力テストではだいたい偏差値70ぐらいであった。しかしA君が都会の超進学校に転校したところ、その高校の実力テストでの偏差値は55になってしまった。A君のお父さんは、これだけ偏差値が下がったのは、転校先の高校の教育が悪いからだと激怒している。あなたがその超進学校でA君の担任教師をしているとするなら、あなたはA君のお父さんにどのようにこの状況を説明するだろう。A君のお父さんが、冷静に話せばわかる人だと仮定して説明してみて下さい。

・センター試験ではなぜ日本史と世界史の平均点が異なっていても、偏差値換算しないのだろう。

・さらに考えてみるなら、入試で国語と英語と社会などの素点を合計するなどといった処置は、数学的にはどのように考えられるのだろうか。国語と英語と社会でのそれぞれの1点の重みは同じなのだろうか。

⇒テスティングの3つの原則

・妥当性 (validity)　そのテストが、本来測りたいと思っている概念（構成概念 construct）を測っているテストであると言えるかどうか。

・信頼性 (reliability)　そのテストを採点するときに、何度採点しても同じ点数が出るようなテストになっているか。また異なる人が採点しても同じ点数が出るようになっているか。

・実施可能性 (practicality)　そのテストは現実的に実施できるか。

(h) 四種類の尺度：名義尺度、順序尺度、間隔尺度、比率尺度の違いを理解しておきましょう。

(7) 名義尺度：好きな野球チームの番号に丸をつけさせるアンケートを行いました。1の広島に丸をしたのが15人だから1x15=15点、2の中日に丸をしたのが11人だから2x11=22点、6の巨人に丸をしたのが10人だから6x10＝60点。したがって巨人が一番人気が高いといえますか？

(8) 順序尺度：ある中学で英語の好き嫌いをアンケートしたところ、1：大嫌いが5人、2：どちらかというと嫌いが13人、3：どちらともいえないが20人、4：どちらかというと好きが18人、5：大好きが2人という結果がでました。この時、「大好き」と「どちらかというと好き」の差は5-4=1として表され、「どちらともいえない」と「どちらかというと嫌い」の差である3-2=1と同じ値をとります。これら二つの好き嫌いの差は質的にも同じだといえますか？また「どちらかというと好き」は4で、「どちらかというと嫌い」は2ですから、前者は後者の二倍ほど英語が好き（あるいは後者は前者の二倍ほど英語が嫌い）といえますか？

あるいはこう考えてみましょう。お肉屋さんに特上、上、並のお肉が並んでいます。特上は並の三倍おいしいのでしょうか。上は並の二倍おいしいのでしょうか。お肉屋さんも、そこまで正確にはお約束しないでしょう。ですが（特殊な好みの違いはここでは考えないことにすると）並よりは上、上よりは特上の方がおいしいという順序は私たちは期待します。このように順序尺度の数字というものは、数字のもつ特性のうちの順序・順番だけを使ったもので、その順序・順番さえ明らかになれば、お肉屋さんのように言葉で表現してもかまいません。順序尺度を統計分析に使うということは、ある意味このようにいいかげんな（？）数字を使っているのだということは覚えておくべきかもしれません。

(9) 間隔尺度：ある英語期末テストでAさんは95点、B君は90点、C君は10点、Dさんは5点を取りました。AさんとB君の学力差は、C君とDさんの学力差と等しいといえますか？

(10) 比率尺度：ある中学生が英検一級の問題を解いてみたところ、やはり中学生には難しすぎて0点を取ってしまいました。この中学生の得点は0点ですから、英語力はゼロだといっていいですか？ この場合の「英語力」の0と、質量や長さの0にはどのような違いがありますか。比率尺度とは絶対的な原点と等間隔な単位を持った尺度と定義できます。

(i) 統計の「厳密さ」について（吉田(1998, pp. 14-16)

社会科学、特に心理学において研究者が本来測定したいと考えている特性は、多くの場合、能力、性格、欲求、動機づけ、思考、感情、価値観といった、人の内面的特性です。しかし、これらの特性は、人間の脳・神経（心）の働きやその表われである人間の行動（の個人差や状況による差異）を説明するために勝手に想定して、いつのまにか、まるで実在するかのように思い込んでしまっているものでしかなく、直接観察して測定することはできません（注：これらの内面的・心理的特性や心理過程のことを、一般に、構成概念とか仮説構成体と呼んでいます）。そのため、私たちは、たとえば、ある課題を完了するのにかかった時間という直接観察できる物理量についてのデータを得ることによって、その課題に関連した能力を各対象がどの程度有しているかといった、能力（の高さ）という量的な内面的特性（これを心理量と呼んでおきます）について間接的に測定（推定）しようとします。また、物理量ではなくても、他者に向ける微笑の頻度からその他者に対する好意度という心理量を測定しようとしたり、何人のクラスメートから「一緒に遊びたい友達だ」と選択されたかという人数によって各児童の人気度という心理量を測定しようとしたりします。しかし、実際にデータとして得られる課題遂行時間や微笑みの頻度や各児童の被選択数といった数値の大きさが、それらを通して測定しようとしている心理量とうまく対応しているとみなしていいのでしょうか。確かに、ある課題を完了するのにかかった時間が短い人ほどその課題に関連した能力が高いと推測することには一般に問題はないと考えられます。しかし、たとえば、ある課題をほとんどの人が5分以内で完了している場合に、その課題を2分で完了させた人の能力の高さと3分で完了させた人の能力の高さの差と、同じ課題を30分かかって完了させた人の能力の高さと31分かかって完了させた人の能力の高さの差が、同程度であるとみなすことにはかなり無理があると考えられます。おそらく、誰もが、後者の1分の差は、その課題を遂行するために必要な能力という心理量においては、前者の1分の差に比べてほとんど意味をもたない（五十歩百歩である）と判断するでしょう。そして、厳密に考えると、得られたデータにおける数値の差が、それを通して測定しようとしている心理量の差に比例しているという意味での望ましい対応関係（図0-2参照）が完全に成立していると断言できるような尺度は存在していないことになってしまうと思います。したがって、なんらかの心理量を測定するためのデータは、すべて、せいぜい順序尺度の変数でしかないことになります。ですから、表面上は間隔尺度または比率尺度の変数だと判断されるものであっても、そこからなんらかの心理量を測定しようとする場合には、尺度の水準を落として順序尺度の変数だとみなし、それに応じた分析を行ったほうが無難だと考えられます。

しかし、実際には、直接観察される量的変数とそれを通して推定される心理量との対応関係に明らかに大きな問題があると考えられる場合でない限り、表面上、間隔尺度ないし比率尺度の変数だと判断されるデータに対しては、それに応じた分析がそのまま適用されています。それは、間隔尺度ないし比率尺度の変数だとみなした方が、表0-1にまとめた各水準のデータに適用できる比較判断の違いからも推察されるように、データがもっている情報を有効に利用できるからです。

また、図0-1で順序尺度の変数に含めた、特定の数量化をすることに問題があると考えられるデータに対しても、多くの場合なんらかの数量化を施し、間隔尺度の変数だとみなした分析が行われています。たとえば、10ページに例としてあげた、ある事柄に対する重要度の認識に関する「非常に重要だと思う」～「まったく重要だと思わない」までの6段階尺度を道いて得られるデータについても、一般に各段階に6～1（または5～0）の等間隔の数値を割り当てて、間隔尺度の変数とみなした分析（たとえば、平均値の算出）が行われます。理由は先ほどと同様です。そこには、少々の問題点に目をつぶっても、そうすることによる利点を重視しようとする前向きな姿勢（？）があるといってもいいかもしれません。しかし、等間隔の数値を割り当てられた各段階の重要度の差が心理的に等しいという保証はどこにもありません。このような意味では、心理学などにおける統計というものは、表面上はさまざまな数学的処理を施して厳密なことをやっているようでも、実際にはかなりラフなことをやっているものなのです。

**参考図書**

副読本として吉田寿夫　(1998)『本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本』北大路書房をお勧めします。この本はまさに題名どおりの良書です。

さらに勉強を深めておきたい人は、森敏昭、吉田寿夫編著(1990)『心理学のためのデータ解析テクニカルブック』北大路書房の第一章と第二章を読んでおくことをお勧めします。

また、統計にあまり強くない人が、間違いなく具体的に統計分析を進めるには、三浦省五監修、前田啓朗、山森光陽編著、磯田貴道、廣森友人著 (2004)『英語教師のための教育データ分析入門』大修館書店が非常に便利です。ぜひ手元に置いておいてください。

理系的素養に乏しい人間として、私が個人的に非常に啓発的な読み物として読んだのが、大村平さんによる一連のシリーズです。彼の著作はロングセラーとして定評があります。とりあえず、『確率のはなし』、『統計のはなし』、『実験計画と分散分析のはなし』、『多変量解析のはなし』(日科技連出版社)を読んでみてください。ただし、こういった本を読む場合は、「速読」や「なんとなくわかる」読みをしないでください。一ページ、一ページ、納得するまでよく考えて、必要ならばノートを取りながら読んでください。そういった地道な勉強が、文系的人間が、少しでも理系的素養を身に付けるために一番必要なことだと思います。なお、これらの本で得られる知識は、テクニカルな知識だけにとどまらず、人生を生きる上でも重要な教養となるものです。